Weekly report

1 Done

1.1 Paper Reading

- ***Disclose More and Risk Less: Privacy Preserving Online Social Network Data Sharing***

The above figure is an example of the attribute inference attack in OSNs. The background knowledge could be the attribute values or connection with actors. Both of them may lead to disclosure. For example, according to #4 and #5 in the attack graph (a), the attacker infers that A and B in the original graph (b) have a high probability of earning $80k–85k annually. According to #6, the attacker knows that around 75% friends of D are in the same age group, which implies that C, as D's friend, is probably around 35 to 39 years old. However, attackers can hardly make accurate inferences on processed graph (c) since only 25% of people in the age group of 20–24 and 40% of engineers have the salary of $80k–85k. Also, for C, after removing the relation between C and D, it is hard to guess C's true age from other attributes.

In this work, the privacy preserving problem is a knapsack-like combinatorial optimization problem. The utility is regarded as profit. The contribution to secret disclosure is regarded as weight. They preserve privacy by removing information directly.

- ***Preventing Private Information Inference Attacks on Social Networks***

This is the first paper that discusses the problem of sanitizing a social network to prevent inference of social network data and then examines the effectiveness of those approaches on a real-world data set. Collective inference is a method of classifying social network data using a combination of node details and connecting links in the social graph. Each of these classifiers consists of three components: a local classifier, a relational classifier, and a collective inference algorithm.

In the wvRN relational classifier takes the neighbors into consideration. Each neighbor is given a weight. The probability of a node being in a class is the weighted mean of the class probabilities of its neighbors.

The privacy preserving process is iteratively generalizing detail type by levels. At the end of each round, they check if the changed graph meets the chosen privacy requirement.

I also read three paper about Bayesian classifiers.

- ***Bayesian Network Classifiers***
- ***Comparing Bayesian network classifiers***
- ***An Improved Naive Bayes Classifier-Based Noise Detection Technique for Classifying User Phone Call Behavior***

Naïve Bayes supposed that the features are independent to each other. So the model is just between features and the target classes.

However, in most cases, features are dependent. Based on that fact, there are some improvements are presented relax the independence assumption.

Tree Augmented Naïve Bayes (TAN) learns a tree structure through mutual information.

The TAN learning procedure is:

1. Take the training set and X \ {c} as input.
2. Call the modified Chow-Liu algorithm. (The original algorithm is modified by replacing every mutual information test I(xi, xj) with a conditional mutual information test I(xi, xj|{c}).
3. Add c as a parent of every xi, where $1 \leqslant i \leqslant n$.
4. Learn the parameters and output the TAN.

BN Augmented Naïve bayes (BAN) extend TAN by allowing the attributes to form an arbitrary graph rather than just a tree.

1. Take the training set and X \ {c} (along with the node ordering) as input.
2. Call a modified CBL1 algorithm – modified by replacing every mutual information test (xi, xj) with a conditional mutual information test I(xi, xj|{c}) , and replacing every conditional mutual information test I(xi, xj|Z) with I(xi, xj|Z {c}) + , where $Z \subset X \setminus \{c\}$.
3. Add c as a parent of every i x where$1 \leqslant i \leqslant n$.
4. Learn the parameters and output the BAN.

General Bayesian Network (GBN) regards the classification nodes as an ordinary node as well.

1. Take the training set and the feature set (along with the node ordering) as input.
2. Call the (unmodified) CBL1 algorithm.
3. Find the Markov blanket of the classification node.

4. Delete all the nodes that are outside the Markov blanket.
5. Learn the parameters and output the GBN.

The last three approaches require O(N2) conditional mutual information tests.

1.2 Group Meeting

- Get ready for presentation.

1.3 Project Idea

- Set the privacy preserving goal as: adversaries can hardly predict sensitive attribute values of individuals. Let n be the amount of the individuals, and a be the number of individuals whose sensitive attribute values are A. We regard that privacy is preserved when adversaries infer that for any individual i, the probability that his/her sensitive value is a, is a/n±e, where e is a user-defined parameter.
- Employ GBN to generate a model includes all attributes.
  - Compute the contribution to useful attributes and sensitive attributes for each attribute values.
- Achieve privacy preservation through generalizing the detail levels until the privacy preserving need is achieved.
- Adjust the information accuracy so that the data quality is consistent.

2 Work Hours

- In all weekdays, I worked during 9:00~11:30, 13:30~5:00 and 18:30~20:30. On weekends, I worked during 11:30~5:00.

3 Progress

Item	Deadline	Current progress	Remark
Graph privacy	-	Change our direction.	-